Four-Injector Variability Modeling of FinFET Predictive Technology Models

Pablo Royer Marisa López-Vallejo Fernando García Redondo Carlos A. López Barrio

***Abstract*—The usual way of modeling variability using threshold voltage shift and drain current amplification is becoming inaccurate as new sources of variability appear in sub-22nm devices. In this work we apply the four-injector approach for variability modeling to the simulation of SRAMs with predictive technology models from 20nm down to 7nm nodes. We show that the SRAMs, designed following ITRS roadmap, present stability metrics higher by at least 20% compared to a classical variability modeling approach. Speed estimation is also pessimistic, whereas leakage is underestimated if sub-threshold slope and DIBL mismatch and their correlations with threshold voltage are not considered.**

I. Introduction

With the reduction of device sizes on future CMOS technologies the mismatch of device parameters is expected to increase and turn to play a major role. Predictive models [1] allow us to foresee the behaviour of the transistors beyond the 22nm node. These models are especially helpful for those circuits that are more sensitive to variations, like SRAMS [2], that occupy large areas of modern-day system on chip. Local mismatch appears more frequently, significantly reducing the reliability of SRAMs if not considered during design.

Commercial technologies usually provide compact models with the variability information included in the model. The parameters in the model card are functions of random variables based on silicon measurements achieving an accuracy suitable for simulations.

Unfortunately this information is not always available, this is the case of predictive technologies where only the nominal device is accessible. Accurate variability simulations are nevertheless required for those models in order to foresee design requirements that would be necessary, such as write-assist techniques [3] or testing new circuits. This has been traditionally bypassed by randomly varying some parameters of the devices. A variation on the dimensions of the transistor, L and W, can be used as an approximation of the devices intrinsic physical parameter mismatch that are at the source of variability.

The two-injector model [5] adds two external power sources to modify the transistor electrical behaviour by tuning its threshold voltage and drain-source current, see Figure 1a. This method presents the advantage of directly dealing with electrical parameters, whose effects in performance are easier to understand, and that are directly measurable from silicon. Notably, the voltage source that models threshold voltage is directly linked to the coefficient AVt, usually reported for a given technology [6], that links the size of the transistors to the magnitude of the mismatch.

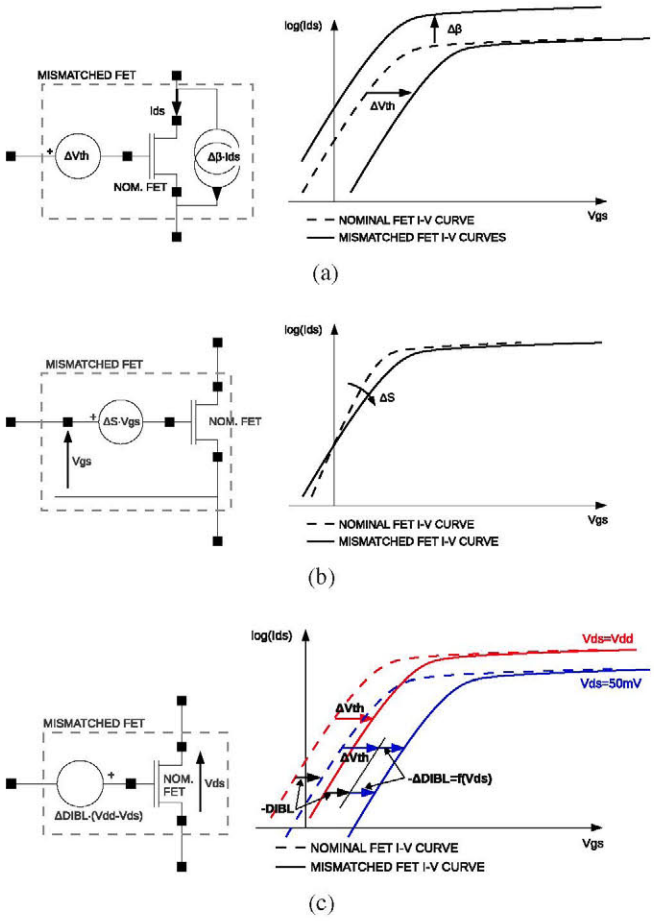

Fig. 1: Injectors used to model (a) threshold voltage and drain-source gain (two-injector approach), (b) sub-threshold slope and (c) DIBL proposed in [4] and their respective effects in the I-V curves of the transistors.

The two-injector model has shown reasonable accuracy for technology nodes down to 45nm, but for upcoming technologies new sources or variability become relevant such as drain induced barrier lowering or sub-threshold slope [7].

Process variation modeling for predictive models is a concern widely considered in the literature. [8] modeled the variability using threshold voltage as a source of variability

to evaluate stability of 9T SRAMs for the 32nm node. [9] performed a full PVT analysis of 6T SRAM cells down to the 7nm node, however their variability modeling only considered two sources of mismatch: EOT and TOXE, which is inaccurate as the technology shrinks.

It was shown in our previous work [4] that a four-injector approach, adding two supplementary power sources to model DIBL and sub-threshold slope mismatches as shown in Figure 1 enhanced the accuracy of Monte-Carlo analysis when simulating the key performance and stability metrics of an SRAM cell, as well as the yield of a whole memory.

In this work we apply the four-injector variability modeling to predictive technology nodes from 20nm down to 7nm using randomly generated injectors, opposed to the injectors generated to fit already known statistical compact models used in [4], and compare the results to those obtained using a two-injector only approach, also randomly generated.

In the next section the methodology used in this work is explained, we first validate the random generation of four-injector sets and we explain how the mismatch figures are scaled across nodes and device sizes and introduce the targets followed during SRAM designs as well as the metrics used. The following section presents the performance and stability results and in the end conclusions are drawn.

II. Methodology

In this section we first validate the method used to randomly generate the injectors for different technology nodes. Then, we explain how we scale the mismatch magnitudes from one node to another based on the transistors areas. Finally we introduce the metrics measured during the variability simulations of the cells and which requirements they have to fulfill based on the ITRS roadmap.

A. Validation of Injectors Generation

In [4] the sets of four injectors were generated to reflect exactly the same variability as the 1000 model cards used as a reference in a one-to-one association. This was useful to validate the accuracy of four-injector variability modeling but its application were limited as the statistical compact model were necessary. We show here how injectors can be randomly generated based on the correlations obtained in [4]. This allows us to apply mismatch to technologies with no variability information available, to scale the mismatch magnitudes based on the device areas to figures reported in the literature as well as generating any arbitrary number of Monte Carlo points.

In order to validate the accuracy of randomly generated injectors we have simulated SRAM cells under variability. A reference of 1000 compact models [10] that include variability were used. Then, 1000 sets of four and two injectors chosen to reflect the same mismatch as the compact models were simulated, This is, for each sample from the compact model, there is one set of two injectors and one set of four injectors that are chosen to match the same variability metrics than the compact model. Finally 10000 Monte-Carlo simulations were run using randomly generated sets of four injectors, there, the mean, standard deviations and correlations of the variability metrics of the generated injectors match those of the compact model, but there is not a one-to-one correspondence.

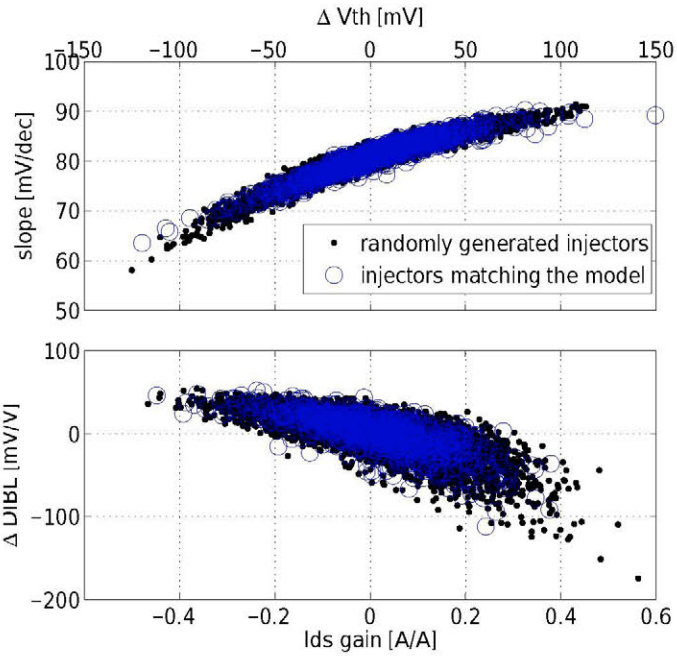

Fig. 2: Threshold voltage, drain-source current gain, DIBL and sub-threshold slope injectors generated following the 1000 model cards, compared to 10000 injectors randomly generated.

Injectors are generated taking into account the fact that the drain to source current, the threshold voltage and the logarithm of the leakage current are linearly correlated on the one hand, as well as the sub-threshold slope and DIBL on the other hand. The sub-threshold slope being directly derived from the leakage current and the threshold voltage.

Figure 2 shows the threshold voltage variation versus sub-threshold slope injectors and drain-source current gain versus DIBL injectors respectively of the 1000 injectors generated from the compact models and 10000 injectors randomly generated using the correlations previously described. A good match between the clouds of the randomly generated injectors and the ones fitting the compact models can be observed.

Also we have calculated the yield of an SRAM array shown in Figure 3 as an additional validation step to compare the accuracy of the different approaches. The already proven enhancement obtained using four injectors instead of two can be seen. In addition it shows a perfect match between randomly generated injectors and the reference four-injector curve, concluding that the good match of the variability metrics observed in the clouds in Figure 2 translates into a good matching of the yield figures.

B. Generation for predictive technology nodes

We use the previously introduced correlations to generate random injectors for other technology nodes, instead of the node for which we already had variability information, and that only was done for validation purposes.

While the correlations are kept constant, the spreads are scaled according to the device areas following Pelgrom's

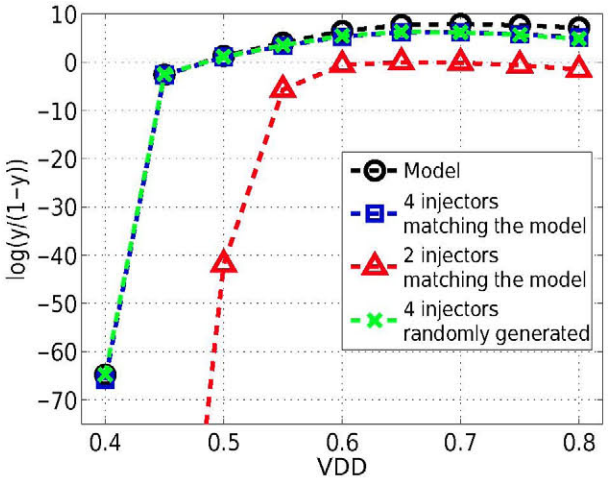

Fig. 3: Yield of an SRAM memory for different supply voltages. The four-injector technique shows a better approximation to the results of the model than only two injectors. The randomly generated sets of four injectors match exactly the results of model generated injectors.

rule [6]:

$$\sigma_{Metric} = \frac{A_{Metric}}{\sqrt{Area}} \tag{1}$$

This rule has been traditionally applied to the threshold voltage mismatch, however it is a generic formula that can be applied to different metrics [11].

In our case, we deal with FinFET devices, whose equivalent area is computed as [12]:

$$Area = L \cdot (2 * H_{fin} + T_{fin}) \cdot N_{fin} \tag{2}$$

where L is the gate length, H_{fin} and T_{fin} are respectively the fin height and width and N_{fin} is the number of fins the transistor is made of.

In addition to scaling the mismatch magnitudes depending on the areas of the devices for each node, we scale them to achieve mismatch figures comparable to those reported in the literature. The coefficient A_{Vt} is chosen to be of $1\ mV\mu m$ which is a value to which current technologies are converging [13], [14]. The other A_{Metric} coefficients for drain-source gain, DIBL and $log(I_{off})$ mismatches are scaled by the same factor than A_{Vt} taking as a reference those measured for the 1000 model cards [10], [4].

C. Memory modeling

In this work we will simulate SRAMs with predictive technology devices corresponding to the 20nm, 14nm, 10nm and 7nm nodes from the Arizona State University [1]. The device dimensions are shown in Table I.

The memories designed in this work follow the International Technology Roadmap for Semiconductors directives [15]. The ITRS SRAM performance targets corresponding to each node of the predictive technology are summarized in Table II.

TABLE I: FinFET dimensions of the Arizona University Predictive Tecnology Models [1] used in this work.

Node	20 nm	14 nm	10 nm	7 nm
Supply Voltage	0.9 V	0.8 V	0.75 V	0.7 V
Gate Length	24 nm	18 nm	14 nm	11 nm
Fin Height	28 nm	23 nm	21 nm	18 nm
Fin Width	15 nm	12 nm	10 nm	7 nm

TABLE II: ITRS targets for SRAMs corresponding to the technology nodes considered in this work.

Node	20 nm	14 nm	10 nm	7 nm
Year	2013	2016	2019	2022
F - half pitch	35 nm	25 nm	18 nm	13 nm
6T Cell Area	$140F^2$			
Delay (HD)	0.80 ns	0.50 ns	0.30 ns	0.30 ns
Delay (HS)	0.15 ns	0.10 ns	0.07 ns	0.07 ns
Leakage Power (HD)	1.5 nW/cell	2 nW/cell	2.5 nW/cell	3 nW/cell
Leakage Power (HS)	1 µW/cell	2 µW/cell	3 µW/cell	5 µW/cell
Metal1 Capacitance	1.9 pF/cm	1.9 pF/cm	1.8 pF/cm	1.6 pF/cm

We will focus on two memory sizes 16kbit for high-speed and 2Mbit for high density. Given the sizes of the memories and the size of the cells ($140F^2$), and the capacitance of the metal lines (C_{M1}) according to ITRS in Table II and assuming the array is designed to have a square aspect ratio, we can calculate the bit-line capacitance as:

$$C_{bit-line} = \sqrt{140F^2} \cdot C_{M1} \tag{3}$$

this bit-line capacitance is one of the main sources of delay and dynamic power consumption in SRAMs.

The limiting delay in SRAMs happens during read operations, that consist on a small cell discharging a large bit-line capacitance in opposition to a write operation, where the bit-line is discharged by the periphery, less subject to variability and able to manage higher currents. Considering that the sense amplifiers would be designed to require a voltage difference of 10% of the supply voltage the delay due to the cells discharging the bit-line will be:

$$Delay_{cell} = V_{sense\ amplifer} \cdot C_{bit-line}/I_{read} \tag{4}$$

where I_{read} is the read current of the SRAM cell.

In other words, assuming that cells account for half of the read delay (the other half is due to the periphery), the read current that the cells will be required to drive is:

$$I_{read} = \frac{V_{sa} \cdot C_{bit-line}}{Delay_{ITRS}/2} \tag{5}$$

where $Delay_{ITRS}$ is the delay requirement set by the ITRS roadmap for each node and each SRAM type (high density or high speed).

Finally, the static power is computed from the leakage power of the cells and the nominal supply voltage of the technology as: $P_{static} = Vdd \cdot I_{leakage}$.

In addition to I_{read} and $I_{leakage}$ performance metrics, that will determine the speed and static power consumption of the cell, read and write static noise margins are simulated to ensure the read stability and write ability of the cells. A cell presenting a negative read static noise margin will not be able to retain its contents during a read access, on the other hand a cell

presenting a negative write static noise margin will not be able to flip its contents during a write access.

All those metrics are simulated under variability obtaining correlated normal distributions for read and write stability metrics [16]. This allows us to infer the failure probability and yield of the memory.

In order to optimize the performance and stability tradeoff of the cells, the number of fins of the transistors can be changed for which we have retained three topologies: 111, 112 and 123, where the numbers represent the number of fins of the pull-ups, the pass-gates and the pull-downs respectively. In addition the nominal threshold voltage of the transistors is allowed to be tuned, this is still possible in FinFET technology, despite a fully-depleted channel, using the gate work function [17].

III. Results

In a first step the nominal metrics of all the possible cells obtained by tuning the threshold voltage of the transistors, and for the three topologies considered were simulated. Cells whose nominal metrics did not meet the stability and performance thresholds introduced before were discarded.

The remaining cells were simulated with mismatch using both four-injector and classical two-injectors methods obtaining their metrics now under variability. The results for those cells are shown in this section.

A. Stability Results

The cells simulated under variability using two and four-injector methods have been filtered so that they meet the speed and static power consumption targets. The cells presenting an optimal read and write stability tradeoff from those passing the speed and power check are shown in Figure 4 for nodes 20nm (top) to 7nm (down) and for the three topologies considered. Each point represents one of the different cells obtained by tuning the threshold voltage of the transistors, given the metrics considered, the variability is already included, a line joins the optimal cells while the others are not represented, fronts for two and four injectors are shown. Those results were obtained for a high density memory, as described in the methodology section, but the same conclusions are drawn for the high speed case.

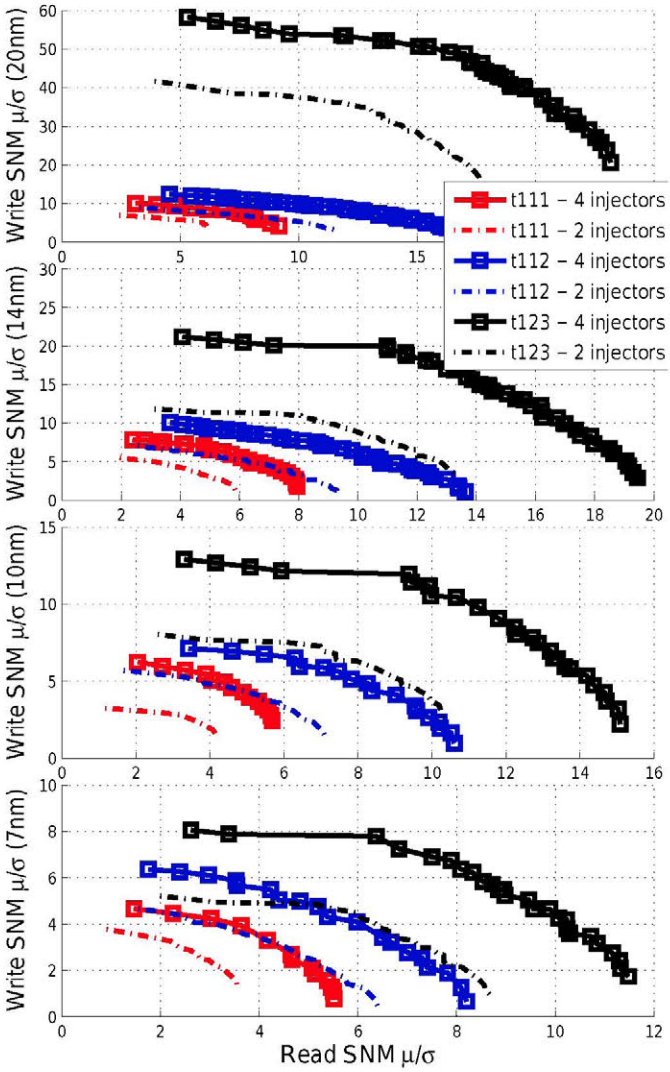

Fig. 4: Read and write stability metrics expressed as mean divided by standard deviation for the nodes and topologies considered.

The metrics shown in the plots are the mean of each metric divided by its standard deviation, which is more representative of cell reliability than just the absolute numbers. As expected given the size of the cells, 123 presents better figures than 112 and than 111. This is due to their larger area, thanks to the higher number of fins, that reduces their mismatch following equations 1 and 2.

In addition we can see that using two-injector method to model variability systematically underestimates the reliability of the cells by at least 20%, this has been calculated as the shortest distance between the curves shown in Figure 4 and with respect to four-injector results. The detailed results for all the nodes and topologies are shown in Table III.

This would lead to wrong conclusions regarding 6T SRAMs viability, if for example we ask cells to have read and write stability metrics above six sigmas, which is necessary to build large arrays with reasonable yields. No cells will achieve this in the 7nm node and only the 123 topology will meet constraints in the 10nm node. A more accurate four-injector approach shows that all topologies are able to achieve six-sigma metrics in the 20 and 14nm nodes, 123 and 112 can make it for 10nm, and it is still conceivable in the 7nm node using the 123 topology.

Some of the numbers shown in Figure 4 might seem very high at first sight, but it has to be noted that this happens for current nodes, that can no longer be considered as predictive, and for a mismatch magnitude that is expected to be achieved in the future as explained in Section II-B. Actually a more realistic scenario would be a mismatch magnitude that would improve together with the technology node, but this is irrelevant as we make the comparisons within the same node.

TABLE III: Detailed underestimation using two-injector method of stability metrics shown in Figure 4 for each technology node and cell topology.

Fin Topology \ Technology Node	20 nm	14 nm	10 nm	7 nm
111	−23.5 %	−26.1 %	−27.6 %	−21.5 %
112	−25.3 %	−26.1 %	−23.1 %	−22.2 %
123	−19.6 %	−29.5 %	−27.9 %	−24.1 %

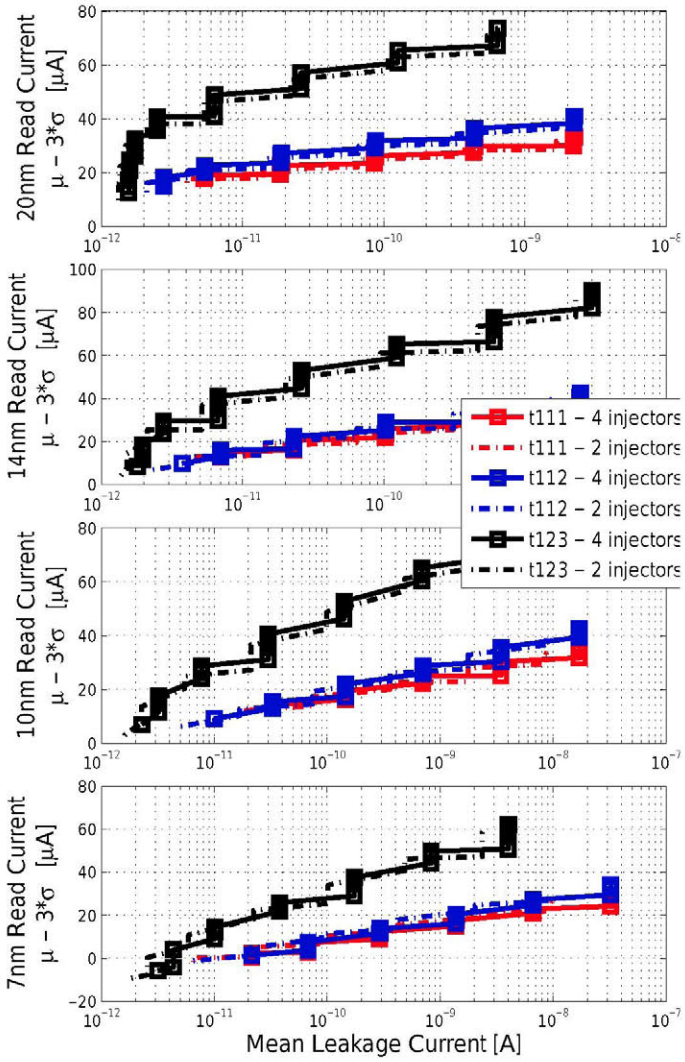

Fig. 5: Static power (limited by leakage) and speed (limited by read current) performance metrics, under variability for all the nodes and topologies considered.

TABLE IV: Detailed underestimation using two-injectors method of the speed metric shown in Figure 5 (vertical axis) for each technology node and cell topology.

Fin Topology \ Technology Node	20 nm	14 nm	10 nm	7 nm
111	−4.3 %	−6.6 %	−6.4 %	−7.7 %
112	−4.9 %	−6.9 %	−6.6 %	−7.3 %
123	−3.7 %	−4.6 %	−5.0 %	−5.8 %

TABLE V: Detailed underestimation using two-injectors method of the mean leakage current shown in Figure 5 (horizontal axis) for each technology node and cell topology.

Fin Topology \ Technology Node	20 nm	14 nm	10 nm	7 nm
111	−24.2 %	−37.5 %	−48.9 %	−69.5 %
112	−24.2 %	−37.5 %	−48.9 %	−69.5 %
123	−15.0 %	−22.8 %	−29.3 %	−41.2 %

B. Performance results

Figure 5 shows the optimal tradeoff of the two main performance metrics: the leakage current, responsible for the static power consumption of the memory, and the read current responsible for the read delay of the memory. As for stabilities, the optimal fronts of the different cells obtained by tuning the threshold voltages are show for two and four injectors.

For the read current we used the mean minus three standard deviations as a metric to take into account variability, as the slowest cell will limit the speed of the whole memory, whereas the mean value of the leakage current is kept, this will determine the static power of the memory even if some cells consume more and other cells consume less.

The plots show that both read and leakage currents are underestimated when only two injectors are used, as a consequence, two-injector model is pessimistic estimating the speed of a memory but optimistic to estimate its static power, with respect to the more accurate four-injector method.

The discrepancy in the read current is mainly due to a higher spread under variability: the standard deviation of the read current is overestimated by 12 to 20% while differences in the mean value are negligible. This leads to a too pessimistic worst case estimation of the read current seen in Figure 5. The detailed results for speed are shown in Table IV.

The additional modeling of sub-threshold slope variability when a four-injector approach is used increases the spread of the leakage current of the transistors and as a consequence, of the SRAM cells. This finally results in an increased mean value of the leakage current for the memory due to the log-normal distribution that this metric follows, making the two-injector method to underestimate the leakage power by 15 to 70% as detailed in Table V.

IV. Conclusions

The random generation of sets of four injectors to model variability is proposed in this work, and used to simulate stability and performance metrics of SRAM memories using

predictive technology models corresponding to nodes 20, 14, 10 and 7nm.

We show that for a nominal power supply stability metrics using only two injectors are systematically underestimated by at least 20%, leading to wrong conclusions regarding the viability of 6T SRAMs. A four-injector variability modeling shows that six-sigma stability metrics is still achievable in the 7nm node using 123 fin topology making large arrays with reasonable yields conceivable.

The reason behind that is an overestimation of the spreads of the metrics, in the same way this affects the read current of the cells whose worst case value is also underestimated by the two-injector method.

As the sub-threshold region is reached the spread in the drain-source current turns to be underestimated by the two-injector approach. The leakage current spread, that was a consequence mainly of threshold voltage mismatch is increased when sub-threshold slope and DIBL variability and their correlation with threshold voltage mismatch.

V. Acknowledgements

This work was funded by CICYT project TOLERA TEC2012-31292 of the Spanish Ministry of Economy and Competitiveness.

References

[1] http://ptm.asu.edu/.

[2] G. Chen, D. Sylvester, D. Blaauw, and T. Mudge, "Yield-Driven Near-Threshold SRAM Design," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 18, no. 11, pp. 1590 –1598, nov. 2010.

[3] V. Chandra, C. Pietrzyk, and R. Aitken, "On the Efficacy of Write-assist Techniques in Low Voltage Nanoscale SRAMs," in *Proceedings of the Conference on Design, Automation and Test in Europe*, ser. DATE '10. 3001 Leuven, Belgium, Belgium: European Design and Automation Association, 2010, pp. 345–350.

[4] P. Royer, P. Zuber, B. Cheng, A. Asenov, and M. Lopez-Vallejo, "Circuit-level modeling of FinFet sub-threshold slope and DIBL mismatch beyond 22nm," in *Simulation of Semiconductor Processes and Devices (SISPAD), 2013 International Conference on*, Sept 2013, pp. 204–207.

[5] P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 6, pp. 1212–1224, 2005.

[6] M. Pelgrom, A. Duinmaijer, and A. Welbers, "Matching properties of MOS transistors," *Solid-State Circuits, IEEE Journal of*, vol. 24, no. 5, pp. 1433 – 1439, oct 1989.

[7] P. Magnone, F. Crupi, A. Mercha, P. Andricciola, H. Tuinhout, and R. J. P. Lander, "FinFET Mismatch in Subthreshold Region: Theory and Experiments," *Electron Devices, IEEE Transactions on*, vol. 57, no. 11, pp. 2848–2856, 2010.

[8] G. K. Reddy, K. Jainwal, J. Singh, and S. Mohanty, "Process variation tolerant 9t sram bitcell design," in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, March 2012, pp. 493–497.

[9] H. Dsilva, J. Pinto, A. Elchidana, and S. Mande, "Variability aware performance evaluation of low power sram cell," in *Quality Electronic Design (ASQED), 2013 5th Asia Symposium on*, Aug 2013, pp. 183–187.

[10] B. Cheng, S. Roy, and A. Asenov, "Statistical Compact Model Parameter Extraction Strategy for Intrinsic Parameter Fluctuation," in *Simulation of Semiconductor Processes and Devices 2007*, T. Grasser and S. Selberherr, Eds. Springer Vienna, 2007, pp. 301–304.

[11] A. Kumar, T. Mizutani, and T. Hiramoto, "Gate length and gate width dependence of drain induced barrier lowering and current-onset voltage variability in bulk and fully depleted silicon-on-insulator metal oxide semiconductor field effect transistors," *Japanese Journal of Applied Physics*, vol. 51, no. 2R, p. 024106, 2012.

[12] H. Dadgour, K. Endo, V. De, and K. Banerjee, "Grain-Orientation Induced Work Function Variation in Nanoscale Metal-Gate Transistors ;Part II: Implications for Process, Device, and Circuit Design," *Electron Devices, IEEE Transactions on*, vol. 57, no. 10, pp. 2515 –2525, oct. 2010.

[13] Q. Zhang, C. Wang, H. Wang, C. Schnabel, D.-G. Park, S. Springer, and E. Leobandung, "Experimental study of gate-first finfet threshold-voltage mismatch," *Electron Devices, IEEE Transactions on*, vol. 61, no. 2, pp. 643–646, Feb 2014.

[14] O. Weber, O. Faynot, F. Andrieu, C. Buj-Dufournet, F. Allain, P. Scheiblin, J. Foucher, N. Daval, D. Lafond, L. Tosti, L. Brevard, O. Rozeau, C. Fenouillet-Beranger, M. Marin, F. Boeuf, D. Delprat, K. Bourdelle, B. Nguyen, and S. Deleonibus, "High immunity to threshold voltage variability in undoped ultra-thin fdsoi mosfets and its physical understanding," in *Electron Devices Meeting, 2008. IEDM 2008. IEEE International*, Dec 2008, pp. 1–4.

[15] http://public.itrs.net/.

[16] H. Park and S. e. a. Song, "Accurate projection of Vccmin by modeling dual slope in FinFET based SRAM, and impact of long term reliability on end of life Vccmin," in *Reliability Physics Symposium (IRPS), 2010 IEEE International*.

[17] M. Jurczak, N. Collaert, A. Veloso, T. Hoffmann, and S. Biesemans, "Review of FinFET technology," in *SOI Conference, 2009 IEEE International*, oct. 2009, pp. 1 –4.